ARMY RESEARCH LABORATORY

ARL

Impact of Machine-Translated Text on Entity and Relationship Extraction

by Mark R Mittrick and John T Richardson

ARL-TN-0649   **December 2014**

NOTICES

Disclaimers

Army Research Laboratory

Aberdeen Proving Ground, MD 21005-5067

ARL-TN-0649   **December 2014**

Impact of Machine-Translated Text on Entity and Relationship Extraction

Mark R Mittrick and John T Richardson
Computational and Information Sciences Directorate, ARL

<table>
<tr><td colspan="2">REPORT DOCUMENTATION PAGE</td><td>Form Approved
OMB No. 0704-0188</td></tr>
</table>

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
December 2014	Final	October 2013–September 2014

4. TITLE AND SUBTITLE

Impact of Machine-Translated Text on Entity and Relationship Extraction

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Mark R Mittrick and John T Richardson

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

US Army Research Laboratory
ATTN: RDRL-CII-C
Aberdeen Proving Ground, MD 21005-5067

8. PERFORMING ORGANIZATION REPORT NUMBER

ARL-TN-0649

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We performed an experiment to study the effects of machine (performed by software) versus manual (performed by a human) translation on the performance of a Small Business Innovation Research text analytics tool. The text analytics in the experiment is Contour, developed by Decisive Analytics Corporation, which automatically builds high-fidelity social networks from text data sets too large to be scrutinized in detail through manual effort. Specifically, we analyzed the ability to extract text entities with the roles of person, location, or organization. The data consists of the translations of many news stories collected from Arabic language websites. There are 5 translations for each story to examine (4 human and 1 machine). The performance of the machine translation Contour results is analyzed against the Contour results of the manual translation.

15. SUBJECT TERMS

foreign language, text analytics, social network analysis, translation, automated, manual, contour

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Mark R Mittrick
a. REPORT	b. ABSTRACT	c. THIS PAGE	UU	28	19b. TELEPHONE NUMBER (Include area code) 410-278-4148
Unclassified	Unclassified	Unclassified			

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

Using social network analysis tools is an important asset in understanding the human terrain in modern operations. The social media that drives these analyses is most often written in a foreign language and requires translation to English before being processed. The translations are performed by humans or, due to the large volume of documents available, by machine (software). These translations are the data used by the social network analysis tools.

It is generally accepted that human-translated documents are superior in content and quality. To test this assumption within the context of social network analysis, we will process a set of translations of Arabic language news articles collected from the web using Contour, a social network analysis tool acquired via a Small Business Innovation Research (SBIR) contract. The set of translations will include 4 separate human translations and 1 machine translation. We will use Contour to process each translation and then collect a list of extracted text entities as the result. Contour is capable of much more in depth analysis, but for our purpose of initial investigation the entity extraction result is sufficient.

Following our assumption, we expect the results from the human translations to be more precise and produce more meaningful entities. In the following sections we will discuss Contour, our method for collecting and analyzing the results, and whether or not our assumption about human versus machine translation is valid in this experiment.

2. Method

The purpose of this experiment is to provide a pilot study that begins an initial investigation into how text derived from machine translation, in contrast to human translation of the same source document, affects the results of a social network analysis tool. To conduct our experiment, we required a social network analysis tool capable of processing text documents and a set of translated documents.

As mentioned previously, the analysis tool in this experiment is Contour (Fig. 1). It is a US Army Research Laboratory Phase II SBIR, being developed by Decisive Analytics Corporation (DAC). Our previous experience with this tool during other experiments, combined with its text analytic capability, made it the best choice for this investigation.
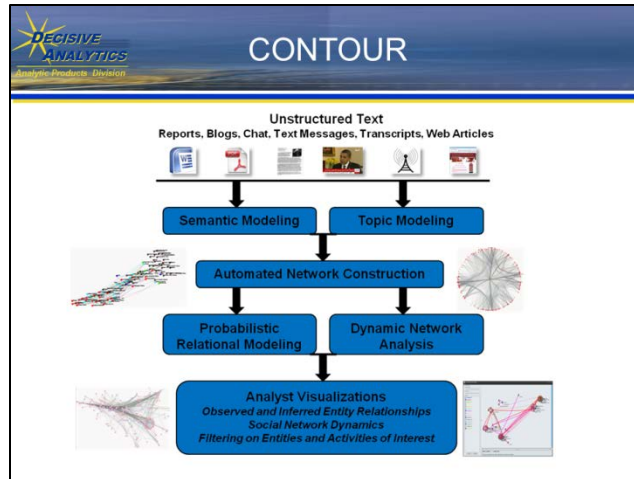
Fig. 1   Contour

Contour provides 4 primary benefits to the tactical Warfighter. First, the system automatically builds high-fidelity social networks from text data sets too large to be scrutinized in detail through manual effort. The mass production of multi-mode social networks from large text corpora will enable analysis on a scale not supported by current techniques. Second, the system is robust to uncertain, incomplete, and conflicting data and does not fail or require user intervention when confronted with the noise of real-world data. Third, automated reasoning capabilities both accelerate the analysis process and conduct more thorough analysis than is possible in a time-constrained manual setting. The result is less time to produce more accurate, actionable intelligence products. Last, it provides visualizations of the complex social network interactions, providing the Warfighter with enhanced situational awareness.[1]

For this experiment, we use the entity extraction capabilities of Contour. Contour uses frame-based semantic modeling software to automatically build detailed network models from unstructured text. Contour imports unstructured text and then maps the text onto an existing ontology of frames at the sentence level, using FrameNet, a structured language model, and through Semantic Role Labeling (SRL). SRL automatically identifies the fundamental concepts expressed by text and maps these ideas into semantic roles.[2] Thus, Contour provided us with a list of extracted entities separated by role (location, people, and organization) for each translation.

After selecting Contour, our next requirement was selecting a translation data set. The set of translated documents needed to be English, small in size, and since the documents may require manual preprocessing, resemble information gathered from social media. As a result, we used The Arabic News Text data set. This data set includes English translations of many Arabic language news articles gathered from the web. In the set each document has 4 human translations and 1 machine translation.

2

Initially, the data was in a single file containing all the documents. As a result, Contour would not process the translations in their original format. Preprocessing of the data was required. The procedure included breaking down the large file into individual files containing a single document and then sorting these files into 5 sets representing each translation. Following these steps, Contour was able to import the data for processing. After each data set was processed, we were able to collect the entities extracted from the text (Fig. 2 shows the results from the machine translation—see the Appendix for all results).

Locations

Iraq
China
Afghanistan
Canada
Cyprus
Madrid
Morocco
Riyadh
Russia
Turkey
Bahrain
Canary Islands
Israel
Saudi Arabia
Taiwan
Washington
Iran
Sanaa
United States
Us

Organizations

European Union
Wto
World Trade Organization
Palestinian National Authority
Saudi Supreme Coordination Council
Army
Taliban
Chinese Central Military Commission
Chinese State Council
Crown
Defence
Foreign Ministry
Defense And Aviation
International Institute
Ministry Of Foreign Affairs
Saudi Coordination Board
Saudi Coordination Council
Second
Ukrainian Defense Ministry
Washington Post

People

Sharon
Abdul Aziz
Abdullah Mereb
Ali
Ibrahim Attia
Abdullah Ii
Al-Mantar
Bo Cheong
Defense Minister Sergei Ivanov
Fahd Bin
Jia Pau
Minister Hani
Minister Kang
Paul Martin
President Abu Mazen
Prime Minister Visiting
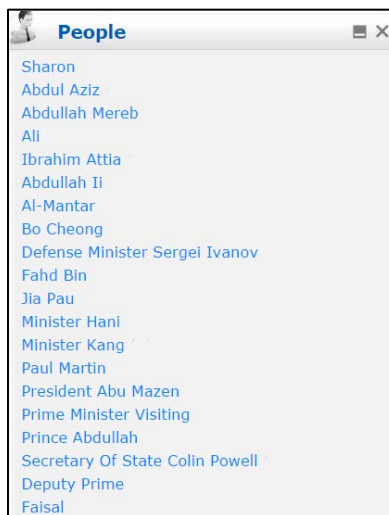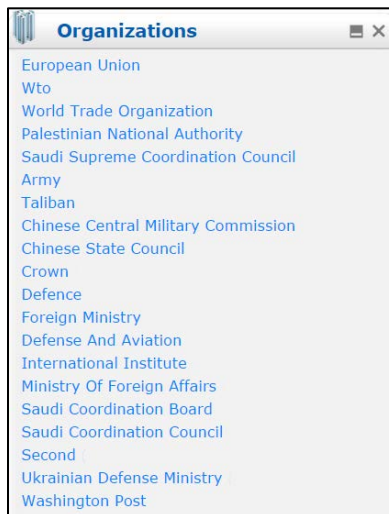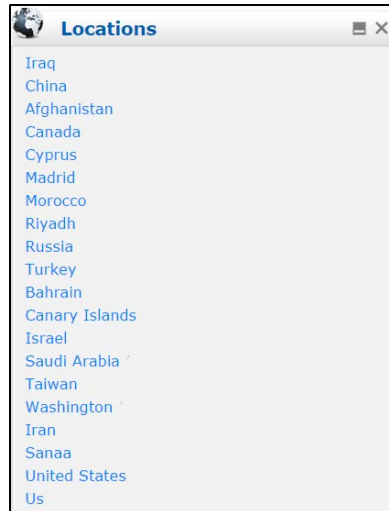Prince Abdullah
Secretary Of State Colin Powell
Deputy Prime
Faisal

Fig. 2   Contour results

3. Results

For this experiment, the data collected included all the text entities extracted from the 5 different translation sets that Contour categorized as people, location, or organization. For each category a master list of entities was created, and the translations from which the entities were extracted were noted (see the Appendix for lists). Table 1 reveals a small part of the master list of entities extracted with the role of Location. Every extracted entity is in the list, and the translations it was extracted from are also recorded.

Table 1   Master list of entities for locations

Locations	Automatic	Manual 1	Manual 2	Manual 3	Manual 4
Afghanistan	X	X	X	X	X
America	. . .	. . .	X	. . .	. . .
Amman	. . .	. . .	X	. . .	. . .
Ankara	. . .	X	. . .	X	X
Baghdad	. . .	. . .	. . .	. . .	X
Bahrain	X	X	X	X	X

To analyze the data in the master lists, all the lists were condensed into a "hit" table. For this analysis, a hit for an entity is defined as that entity being extracted from a translated document. For example, in Table 1 each "X" in an entity's row is a hit. The hit table is represented in Table 2. When creating this table, each text entity is considered to have 0, 1, 2, 3, or 4 hits, depending on how many of the manual translations—which we are using as ground truth—were a source for this entity. A hit from the machine translation is also tallied in the table but is not used to categorize the entity. In terms of hits, a 4 is the best case since the entity was extracted from all 4 manual translations, which is our best indicator that the entity is correct. On the contrary, a zero is the worst case since the entity was only extracted from the machine translation. For example, looking at Table 2, the row for one hit shows that 76 entities were extracted from exactly one manual translation. In addition, 15 of these entities were also extracted from the machine translation.

Table 2   Hit table

Hits	Manual	Machine	Machine Recall (%)	Total Text Entities (%)
0	0	18	NA	NA
1	76	15	20	58
2	22	5	23	17
3	13	5	38	10
4	20	18	90	15

If we only consider the best case of 4 hits where the entities of the manual translations are in complete agreement with each other, we see that Contour was able to recall 90% of the 4 hit entities. The recall rate decreases drastically for the lower hit entities. Furthermore, calculating the precision (Table 3) of extracting 4 hit entities from the manual and machine translation they are very similar, 33% and 30%, respectively. Consequently, Contour was able to extract many more entities from the manual translations, but the precision in extracting the most reliable data was similar as when analyzing the machine translation. In other words, the machine translation result agrees with the overall manual translation results as much as the individual manual translations agree with each other.

Table 3   Four hit precision

Translation Type	Precision (%)
Manual	33
Machine	30

4.   Conclusions

Our original assumption was that the text analytic tool will provide more precise results using the manually translated data than when using the machine-translated data. This experiment has shown us that this is not necessarily the case. The combination of machine-translated data and a robust text analytics tool was able to compare favorably to the same analysis using manually translated data, in as much as the manual translations agreed with each other. There were enough differences across the set of manually translated documents that limited the precision of extracting the exact same entities from each of them, bringing the overall performance in line with extractions from a machine translation. This pilot study is a first glance at examining how reliable machine translations can be in social network analysis. Further investigations using larger and more defined data sets plus leveraging several other analysis tools are recommended.

5. References

1. Army Research Laboratory (US) Fact Sheet. SNARE: Social network realization and explotation. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2013 Aug.

2. Kase SE, Dumer J. Accelerating exploitation of low-grade intelligence through semantic text processing of social media. Paper presented at: ICCRTS 2013. Proceedings of the 18th International Command & Control Research & Technology Symposium; 2013 Jun 19–21; Alexandria, VA.

INTENTIONALLY LEFT BLANK.

Appendix. Data Tables

Table A-1   Locations

Locations	Automatic	Manual 1	Manual 2	Manual 3	Manual 4
Afghanistan	X	X	X	X	X
America			X		
Amman			X		
Ankara		X		X	X
Baghdad					X
Bahrain	X	X	X	X	X
Beijing		X		X	X
Belgium		X			X
Cairo		X			
Canada	X	X	X	X	X
Canary Islands	X	X	X	X	X
China	X	X	X	X	X
Cyprus	X	X	X	X	X
Ebla		X	X	X	
Hong Kong					X
Iran	X	X	X	X	X
Iraq	X	X	X	X	X
Israel	X	X		X	
Jordan				X	
Madrid	X	X	X		X
Manama		X			
Morocco	X		X		
Riyadh	X	X	X	X	X
Russia	X	X	X	X	X
Sanaa	X				
Saudi Arabia	X		X	X	
Syria					X
Taiwan	X		X	X	
The Gulf			X		X
Turkey	X	X	X	X	X
United States	X	X	X	X	X
US	X				
Washington	X			X	

10

Table A-2   Organizations

Organizations	Automatic	Manual 1	Manual 2	Manual 3	Manual 4
Allied		X			
Al-Qaeda				X	
American Army				X	X
Army	X	X	X	X	X
As World Trade Organization			X		
Chinese Central Military Commission	X	X	X	X	X
Chinese Premier		X		X	X
Chinese State Council	X		X		
Cooperation Council		X		X	
Coordinating Council				X	
Coordination Council		X	X		X
Crown	X				X
Defence	X				
Defense and Aviation	X	X	X		
Doha Development		X	X	X	X
Enzo		X	X		X
European Union	X	X	X	X	X
Foreign Affairs				X	
Foreign Ministry	X				X
Foreign Ministry Of Powell			X		
France-Presse		X	X		
Gulf Cooperation Council					X
International Institute	X				X
Islamic Group				X	X
Islamic Group Of Moroccan Fighters		X		X	X
Jordanian Foreign Affairs				X	
Jordanian News Agency				X	
Kazakh				X	X
London Based International Institute		X			
Ministry Of Foreign Affairs	X				
Palestinian Authority		X			
Palestinian National Authority	X		X	X	X
Saudi Coordination Board	X				
Saudi Coordination Council	X				
Saudi News Agency			X	X	
Saudi Press Agency			X		
Saudi Supreme Coordination Council	X				
Saudi Yemeni Coordinating Council				X	
Saudi Yemeni Coordination Council		X			X
Saudi Yemeni Joint Coordination Council			X		
Second	X		X		
Taliban	X	X	X	X	X
The Council					X
Ukrainian Defense Ministry	X				
Ukrainian Foreign Ministry		X			
Underline Sustainable Development			X		
Washington Hopes Breakdown In Communications		X			
Washington Hopes Communications			X		
Washington Post	X	X			
World Trade Organization	X	X	X	X	X
WTO	X				

Table A-3   People

People	Automatic	Manual 1	Manual 2	Manual 3	Manual 4
Abdual Latif					X
Abdul Aziz	X	X	X		X
Abdullah Mereb	X				
Abdullah Murib					X
Adbullah Ii	X	X		X	X
Al-Aqsa			X	X	
Al-Faisal		X			
Al-Faysal				X	
Al-Haski		X			
Ali	X				
Ali Fahimi					X
Al-Islamiya			X		
Al-Jamaa			X		
Al-Malqi				X	X
Al-Manama				X	
Al-Mantar	X		X		
Al-Mintar				X	
Al-Mulqi		X			
Al-Muqatila			X		
Al-Qaida			X		
Al-Qassam			X	X	
B.C				X	
B.C. Found		X			
B.C. Religious					X
Bo Cheong	X				
Christian Chenot				X	
Christian Chesnot		X			
Commander-In-Chief					X
Defense Minister Cao Gangchuan		X		X	X
Defense Minister Olexander Kuzmuk			X		
Defense Minister Sergei Ivanov	X	X		X	X
Deputy Prime	X		X		
Elba					X
Fahd Bin	X				X
Faisal	X	X			
General David Barno		X	X	X	X
Georges Malbrunot		X			
Ibrahim Atia					X
Ibrahim Attia	X				
Inspector General Prince Sultan Ibn			X		
Isa Al Khalifa				X	
Jamal				X	X
Jia Pau	X				
Leader Of				X	X
Leonid Kotchma		X	X		
Mark Mccann		X			
Mattie		X	X		
Meets			X		
Minister Hani	X	X			
Minister Hani Mulki			X		
Minister Kang	X				
Paul Martin	X				
Petra				X	
President Abu Mazen	X		X		
Prime Minister Abd				X	X
Prime Minister Abdul Qader Bajamal			X		
Prime Minister Paul Martin		X			
Prime Minister Visiting	X				
Prince Abdullah	X				
Prince Abdullah Bin					X
Prince Saud		X			
Secretary Of State Colin Powell	X			X	
Sharon	X	X	X	X	X
Theo Van			X	X	X
Wen Jiabao		X			

Table A-4   Automatic translation, locations

Table A-5   Automatic translation, organizations

Table A-6   Automatic translation, people
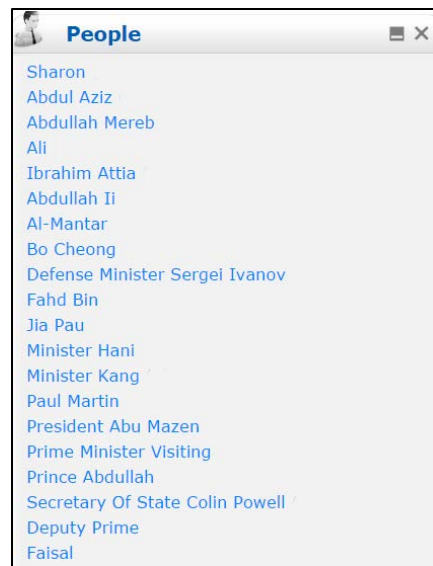

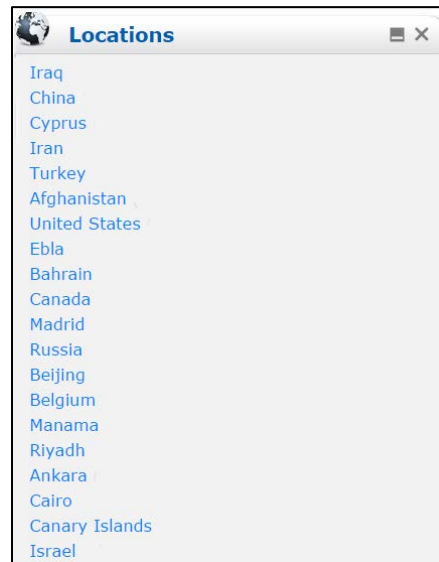Table A-7   Manual 1 translation, locations

Table A-8   Manual 1 translation, organizations

Organizations

European Union
Army
World Trade Organization
Islamic Group Of Moroccan Fighters
Saudi-Yemeni Coordination Council
Cooperation Council
London-Based International Institute
Taliban
Chinese Premier
Enzo
Palestinian Authority
Ukrainian Foreign Ministry
Washington Post
Allied
Chinese Central Military Commission
Coordination Council
Doha Development
Washington Hopes Breakdown In Communications
Defense And Aviation
France-Presse

Table A-9   Manual 1 translation, people

People

Al-Mulqi
Abdullah Ii
Al-Faisal
Mattie
Minister Hani
Sharon
Faisal
Prime Minister Paul Martin
Prince Saud
Wen Jiabao
Abdul Aziz
Al-Haski
B.C. Found
Christian Chesnot
Defense Minister Cao Gangchuan
Defense Minister Sergei Ivanov
General David Barno
Georges Malbrunot
Leonid Kotchma
Mark Mccann
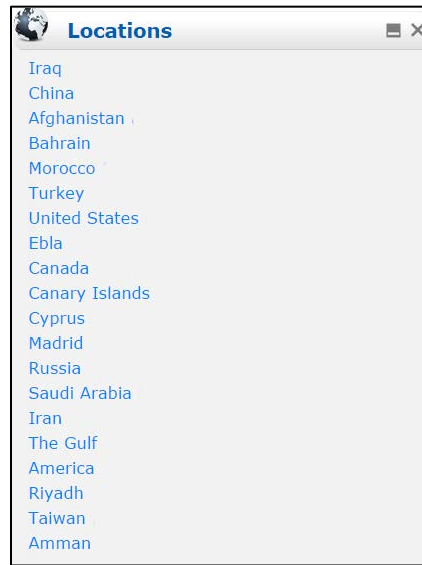
15

Table A-10   Manual 2 translation, locations

Locations

Iraq
China
Afghanistan
Bahrain
Morocco
Turkey
United States
Ebla
Canada
Canary Islands
Cyprus
Madrid
Russia
Saudi Arabia
Iran
The Gulf
America
Riyadh
Taiwan
Amman

Table A-11   Manual 2 translation, organizations

Organizations

Army
World Trade Organization
European Union
Saudi Yemeni Joint Coordination Council
Palestinian National Authority
Taliban
Coordination Council
Defense And Aviation
Saudi News Agency
Second
Underline Sustainable Development
As World Trade Organization
Chinese Central Military Commission
Chinese State Council
Enzo
Foreign Ministry Of Powell
France-Presse
Saudi Press Agency
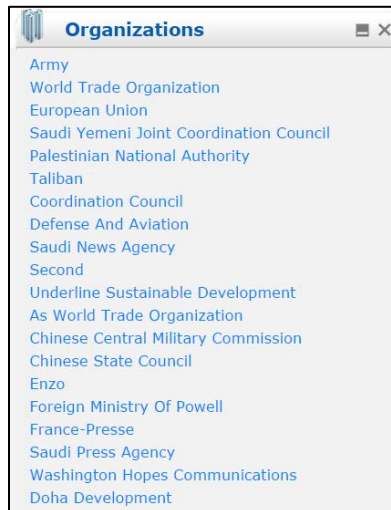Washington Hopes Communications
Doha Development

16

Table A-12   Manual 2 translation, people

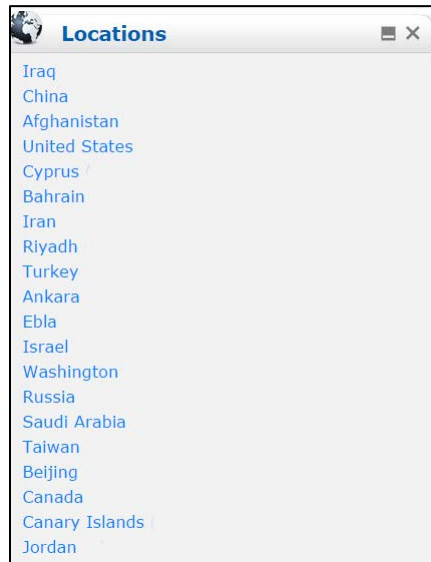Table A-13   Manual 3 translation, locations

Table A-14   Manual 3 translation, organizations

Organizations

European Union
Army
Saudi-Yemeni Coordinating Council
World Trade Organization
Foreign Affairs
Islamic Group Of Moroccan Fighters
Palestinian National Authority
Saudi News Agency
Taliban
Jordanian Foreign Affairs
Doha Development
Islamic Group
Jordanian News Agency
Kazakh
Al-Qaeda
American Army
Chinese Central Military Commission
Chinese Premier
Cooperation Council
Coordinating Council

Table A-15   Manual 3 translation, people

People

Al-Malqi
Sharon
General David Barno
Theo Van
Abdullah Ii
B.C
Isa Al Khalifa
Al-Faysal
Al-Mintar
Jamal
Leader Of
Petra
Prime Minister Abd
Secretary Of State Colin Powell
Al-Aqsa
Al-Manama
Al-Qassam
Christian Chenot
Defense Minister Cao Gangchuan
Defense Minister Sergei Ivanov
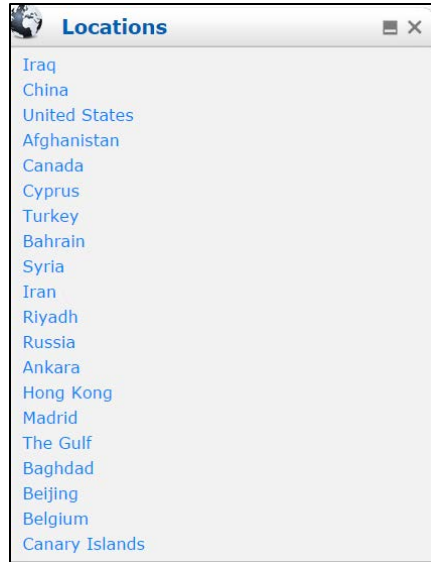
18

Table A-16   Manual 4 translation, locations

Locations

Iraq
China
United States
Afghanistan
Canada
Cyprus
Turkey
Bahrain
Syria
Iran
Riyadh
Russia
Ankara
Hong Kong
Madrid
The Gulf
Baghdad
Beijing
Belgium
Canary Islands

Table A-17   Manual 4 translation, organizations

Organizations

Army
European Union
World Trade Organization
Taliban
Palestinian National Authority
Saudi-Yemeni Coordination Council
Doha Development
Islamic Group
American Army
Crown
Islamic Group Of Moroccan Fighters
Kazakh
The Council
Chinese Central Military Commission
Chinese Premier
Coordination Council
Enzo
Foreign Ministry
Gulf Cooperation Council
International Institute

Table A-18   Manual 4 translation, people

People

Al-Malqi
Abdul Aziz
Abdullah Murib
Ali Fahimi
Ibrahim Atia
Sharon
Theo Van
Fahd Bin
General David Barno
Jamal
Prime Minister Abd
Prince Abdullah Bin
Abdul Latif
Abdullah Ii
B.C. Religious
Commander-In-Chief
Defense Minister Cao Gangchuan
Defense Minister Sergei Ivanov
Elba
Leader Of

Table A-19   Recall tables

Locations

Hits	Manual	Machine	Machine Recall
0	0	2	NA
1	10	2	20%
2	5	3	60%
3	4	1	25%
4	12	12	100%

Organizations

Hits	Manual	Machine	Machine Recall
0	0	7	NA
1	25	6	24%
2	8	1	13%
3	5	1	20%
4	6	5	83%

People

Hits	Manual	Machine	Machine Recall
0	0	9	NA
1	41	7	17%
2	9	1	11%
3	4	3	75%
4	2	1	50%

INTENTIONALLY LEFT BLANK.